A First Step Towards Content Protecting Plagiarism Detection

Cornelius Ihle
cornelius.ihle@daimler.com
Daimler AG & Univ. of Wuppertal
Stuttgart, Germany

Moritz Schubotz
moritz.schubotz@fiz-karlsruhe.de
FIZ Karlsruhe
Berlin, Germany

Norman Meuschke
meuschke@uni-wuppertal.de
Universities of Wuppertal & Konstanz
Wuppertal & Konstanz Germany

Bela Gipp
gipp@uni-wuppertal.de
Universities of Wuppertal & Konstanz
Wuppertal & Konstanz Germany

ABSTRACT

Plagiarism detection systems are essential tools for safeguarding academic and educational integrity. However, today's systems require disclosing the full content of the input documents and the document collection to which the input documents are compared. Moreover, the systems are centralized and under the control of individual, typically commercial providers. This situation raises procedural and legal concerns regarding the confidentiality of sensitive data, which can limit or prohibit the use of plagiarism detection services. To eliminate these weaknesses of current systems, we seek to devise a plagiarism detection approach that does not require a centralized provider nor exposing any content as cleartext. This paper presents the initial results of our research. Specifically, we employ Private Set Intersection to devise a content-protecting variant of the citation-based detection method Bibliographic Coupling implemented in our plagiarism detection system HyPlag. Our evaluation shows that the content-protecting method achieves the same detection effectiveness as the original method while making common attacks to disclose the protected content practically infeasible. Our future work will extend this successful proof-of-concept by devising plagiarism detection methods that can analyze the entire content of documents without disclosing it as cleartext.

CCS CONCEPTS

• **Information systems → Near-duplicate and plagiarism detection**; *Hashed file organization.*

KEYWORDS

Similarity Detection; Plagiarism Detection; Private Computation

ACM Reference Format:
Cornelius Ihle, Moritz Schubotz, Norman Meuschke, and Bela Gipp. 2018. A First Step Towards Content Protecting Plagiarism Detection. In *JCDL '20: ACM/IEEE Joint Conference on Digital Libraries, June 19–23, 2020, Wuhan, Hubei, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
JCDL '20, June 19–23, 2020, Wuhan, Hubei, China
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Plagiarism, i.e., the unacknowledged reuse of ideas or content, is a severe form of academic misconduct. Today, educational and research institutions, academic publishers, and funding agencies increasingly rely on plagiarism detection systems (PDS) to identify plagiarized content [21]. Typical PDS require users to submit input documents, which the systems then compare to a large, typically proprietary database of documents. The systems retrieve comparison documents with similar content as the input document and highlight the similar content to support user inspection.

Researchers and practitioners have criticized PDS for their poor detection accuracy and opaque computations [21], as well as their centralized and nontransparent data management [20, p. 72ff.]. In past research, we have addressed the first issue. We improved detection rates for heavily disguised instances of academic plagiarism by integrating the analysis of text-independent content elements, such as academic citations, images, and mathematical content with text-based detection methods into the hybrid plagiarism detection system HyPlag [14]. In this paper, we focus on the second issue.

Disclosing the full content of input and comparison documents to a central service provider whose detection methods and data protection measures are nontransparent, inherently raises concerns regarding the security and confidentiality of sensitive data, e.g., in unpublished research grant proposals or research theses compiled in cooperation with companies. The mere disclosure of such content to a third party can violate non-disclosure agreements, as well as data protection and copyright laws [20, p. 73f.]. Such legal concerns can limit the use of plagiarism detection systems. The general risk of data breaches further aggravates the problem. In Germany, many universities, therefore, prohibit the use of PDS entirely.

As we described in our vision paper [7], we seek to address the weaknesses of current PDS by devising a blockchain-backed decentralized approach to plagiarism detection (PD) that does not require disclosing any content as cleartext.

As a first step towards this vision, this paper reports on devising a content-protecting variant of the citation-based similarity measure Bibliographic Coupling (BC) [10] implemented in our PDS HyPlag [14]. We present an approach to securely mask bibliographic references and an adaption of the Private Set Intersection (PSI) approach to compute the bibliographic coupling strength (BCS) of papers without revealing the cleartext of the references.

2 RELATED WORK

Our research shall enable plagiarism detection systems to identify similar content in documents without disclosing the content. Data security research has yielded two approaches to protect content while allowing the identification of the cleartext: Revertible functions (encryption) and lossy one-way functions (hashing).

Encryption is conceptually less secure than hashing because the encrypted content contains the full information of the cleartext. A malicious party can gain access to the cleartext by obtaining the decryption key. Additionally, encryption methods considered secure today can become vulnerable in the future due to undiscovered flaws or increases in computing power [3].

Hash functions are lossy one-way functions that map arbitrary-sized cleartext to a fixed-sized value (hash). Due to the lossy mapping, the cleartext cannot be recomputed from the hash. The only option for a malicious party to ascertain the correspondence of a hash and the cleartext is to guess the possible cleartext, compute its hash, and compare it to the hashes disclosed for a document. This approach, known as a *preimage attack*, is feasible if the set of hash inputs is finite and known [18]. Privacy-focused messaging applications like Signal face a similar problem called private contact discovery due to the finite set of phone numbers [11]. Signal solved this issue by conducting PSI in a secured part of the CPU [12]. This solution does not apply to our distributed detection use case, as it still requires trust in the hardware underlying the service.

Researchers predominantly employed hashing for the detection of document similarity without revealing the documents' content.

The work of Unger et al. [19] is most related to ours since it also protects the content of documents in the plagiarism detection process. The authors' approach relies on a central authority (root) that manages the access of nodes to the distributed system in general and to specific content. The nodes represent documents as word chunks, which they process using a collision-resistant hash function with a globally shared salt. The computed hashes are added to a count-min sketch [4] for tracking the frequency of word chunks in the document. The count-min sketches are shared with the root and can be queried by the nodes. The privacy of all communication within the system is secured using the TLS protocol, with the root acting as a certificate authority. While the approach greatly improves the confidentiality of content compared to traditional PDS, the count-min sketches are vulnerable to dictionary attacks. Furthermore, the root receives meta-data about documents, which potentially includes the documents' subject matter and information on the authors' writing style.

Murugesan et al. [16] proposed the use of bloom filters in combination with hashes to protect the semantic meaning of content but maintain knowledge about the content's composition to perform similarity detection tasks. Bloom-filters are conceptually related to count-min sketches. Garbled Boom Filter proposed by Murugesan et al. [16] track the frequency of content chunks in a document using a fixed-sized map while filling the empty positions with noise.

A drawback that all of the aforementioned hash-based approaches share is their vulnerability to preimage attacks. Furthermore, the approaches rely on a central authority.

Secure Multi-Party Computation (SMPC) [6] describes methods that overcome the need for a central authority. In SMPC, parties jointly compute a function over inputs, which the parties keep private. Most SMPC protocols, however, require translating all computations to binary circuits [17]. Employing SMPC for plagiarism detection would thus require new implementations of PD methods.

Many PD tasks represent an exchange of data between two instead of n parties, hence do not require the application of elaborate SMPC protocols. Instead, the tasks can be solved using the *Private Set Intersection (PSI)* [2] approach. PSI allows two parties to compare private versions of their sets of data without revealing information to third parties. Hashing is the core of PSI. Many PD tasks exclusively require ascertaining the existence of identical features in documents of two parties. Hence, we consider PSI as promising for developing content-protecting versions of PD methods.

3 METHOD

In this paper, we consider bibliographic references as the only content to be compared confidentially. The bibliographic coupling algorithm considers the sets of references in the input and comparison document R_d and R'_d to compute the document similarity score bibliographic coupling strength (s_{BC}) as

$$s_{\mathrm{BC}}\left(R_d, R'_d\right) = \frac{\left|R_d \cap R'_d\right|}{\left|R_d \cup R'_d\right|} = \frac{\left|R_d \cap R'_d\right|}{|R_d| + |R_{d'}| - \left|R_d \cap R'_d\right|}. \quad (1)$$

Employing simple hashing to protect the confidentiality of bibliographic references is prone to preimage attacks as the number of published papers, and hence the number of possible references is finite. An attacker could acquire the metadata of most or all references, pre-compute their hashes and compare the pre-computed hashes of arbitrary references to the hashes of a protected document to deduce the professional context of the document.

To prevent preimage attacks for our use case, we hash combinations of k references instead of single references and only disclose the resulting set of combined hashes to the detection service.

Without loss of generality, we assume that a preprocessing of references has been completed before forming the subsets. We further assume that the processing step i) eliminated any duplicates in the reference lists of individual documents, ii) disambiguated all references in the collection, iii) stored the disambiguated references as a hashable data structure, and iv) excluded documents that contain less than k references from the similarity computation.

We form all k-combinations. For example, if a document contains the set of three references $\{a, b, c\}$ and we seek to form subsets of cardinality $k = 2$, we would form, e.g., the subsets $\{a, b\}$, $\{a, c\}$, and $\{b, c\}$ but not $\{a, a\}$. Formally, we form the reference subsets

$$\mathcal{P}_k(R_d) = \left\{r \subseteq R_d \middle| \, |U| = k\right\} \quad (2)$$

with cardinality $|\mathcal{P}_k(R_d)| = \binom{|R_d|}{k}$. Growing the set of possible hashes by a power of k increases the cost of a preimage attack but also the complexity of the detection process. Moreover, document pairs must contain at least k common references to exhibit a similarity that the content-protecting BC method can detect.

To mask the cleartext of references, we compute the set of hashes

$$H_d = \left\{H(r) \middle| r \in \mathcal{P}_k(R_d)\right\}. \quad (3)$$

Here $H(r) = \sum_{i=1}^{k} H(r_i)$ denotes the hash function over the subset of references $r \subseteq R_d$. For the case $k = 1$, $H(r)$ yields the hashes of the individual references, i.e., $H_d = \{H(r_1), H(r_2), \ldots, H(r_n)\}$.

The detection server performs a private set intersection of the hashes from the input document and the hashes from previously submitted documents H'_d to compute the private BCS as

$$s_{\text{PBC}}(H_d, H_{d'}) = \frac{|H_d \cap H_{d'}|}{|H_d \cup H_{d'}|}. \quad (4)$$

Similarly, one can derive $s_{\text{BC}}(H_d, H_{d'})$ via

$$s_{\text{BC}}(H_d, H_{d'}) = \frac{\Im_k(H_d \cap H_{d'})}{\Im_k(H_d) + \Im_k(H_{d'}) - \Im_k(|H_d \cap H_{d'})}, \quad (5)$$

where $\Im_k(j)$ is the numeric solution for $j = \binom{\Im_k(j)}{k}$. For example, for $k = 2$ one can derive $\Im_k(j) = \frac{1}{2}(1 + \sqrt{8j+1})$.

After comparing a private document's hashes to the hashes of the corpus, we retrieve potential source documents by ranking all comparison documents in descending order of their maximum s_{PBC} and filter for matches exclusively occurring in one document pair.

4 EXPERIMENTS

Our experiments analyze the effectiveness, consumption of computational resources, and resistance to preimage attacks of our content-protecting version of the BC algorithm.

4.1 Experimental Setup

We conducted our experiments on a dataset of 105,120 arXiv documents, into which we embedded 10 cases of confirmed plagiarism, each consisting of the plagiarized document and one source document. We used the same dataset in a previous work [15]. We excluded documents without processable reference data and documents with more than 150 references. The final dataset contained 92,082 documents and 1,726,359 unique bibliographic references.

In a preprocessing step, we used the open-source software GROBID[1] to convert all documents into the uniform TEI-format[2]. TEI employs XML to structure the documents' content and allows for easy extraction of the bibliographic references.

Initial tests showed that the title is the reference field with the highest probability of being present in the reference string. Therefore, we used the normalized title field for the hashing. To decide on a hash function to use, we counted the number of hash collisions resulting from using Adler32, SHA1, and SHA256 for hashing reference subsets of size 3 in 5,000 documents. Only Adler32 yielded hash collisions (1,320), i.e., mappings of different inputs to the same hash, due to the comparably smaller size of its hashes (32-bits). We thus chose SHA1 as it offers sufficient collision resistance (2^{80}) and is faster to compute than SHA256. Collisions would cause overmatching and unnecessary effort for human reviewers.

4.2 Results

Effectiveness. To compare the effectiveness of BC-PSI to the original BC method, we computed s_{PBC} and s_{BC} for all ten test cases in our dataset using subset sizes of 1, 2, and 3 respectively. We found that for both $k = 2$ and $k = 3$, the similarity scores computed by BC-PSI and BC were equal for all test cases. This result shows that BC-PSI detects identical references equally well as BC.

Table 1: Hash generation for 92,082 documents.

k-Tuple	Hashes	Time in sec	Size in GB
k = 1	1,726,359	21	0.185
k = 2	45,951,328	31	2.5
k = 3	1,848,313,500	258	126

Table 2: Detection against 1,000 documents

k-Tuple	Hashes	Ratio in 1/2/3 docs	Time in ms
k = 1	22,658	.86/.07/.03	98
k = 2	357,765	.98/.02/.001	103
k = 3	5,250,076	.99/.01/.0	118

Resource Consumption. To assess the computational effort of our content-protecting BC-PSI method, we analyzed the computation time and storage required for computing s_{PBC} depending on the size of k. We divided the analysis into two steps.

In the first step, we assessed the time and storage required for computing and storing the hashed reference subsets. Table 1 shows the results for analyzing the entire dataset of 92,082 documents using different sizes of k. The results show that the exponential growth of the space required for storing hashed reference subsets is the limiting factor for using larger subset sizes.

In the second step, we assessed the time required for performing the private set intersection of the hashed reference subsets depending on k. For this analysis, we only used 1,000 documents from our dataset. Table 2 shows the number of all hashes and the fraction of those hashes that occur in one, two, and three documents, respectively. The table also shows the time required for applying BC-PSI to compute the bibliographic coupling strength of the input document with a comparison document. For increasing values of k, the number of hashes occurring in more than one document decays rapidly. For $k = 1$, 86% of the references occur in one document only. Combinations of three references are unique in 99.3% of the cases. The increase in required computation time is almost constant in the number of hashes due to the use of effective indexes.

Resistance to Preimage Attacks. To motivate the resistance of BC-PSI to preimage attacks, we estimate the computing time required for such an attack. As explained in Section 2, a preimage attack requires knowing the possible hash inputs, i.e., in our case, the possible references cited in academic documents.

Jinha estimated that the number of published journal articles surpassed 50 million in 2009 [8]. Other studies consistently found that the annual growth rate of published journal articles is approx. 3% [1, 9]. By extrapolation, we estimate that 67 million journal articles existed at the end of 2019[3]. This number underestimates the number of possible references as it does not consider, e.g., conferences papers, books, or preprints. However, lacking reliable estimates on the number of sources other than journal articles, we use 67 million as a lower bound estimate on the number of possible references.

[1] https://github.com/kermitt2/grobid
[2] https://tei-c.org/
[3] $50 \times 10^6 * 1.03^{10} = 67.19 \times 10^6$

Given the set of possible references contains 67 million elements, a preimage attack on a single input document for $k = 1$ requires the computation of $\binom{67\times10^6}{1} = 67 \times 10^6$ hashes, leading to $O(n^k)$. Assuming a computation time of 1ms per hash results in an overall computation time of $67 \times 10^3 s \approx 18.61h$. Analogously, for $k = 2$, computing $\binom{67\times10^6}{2} \approx 2244 \times 10^{12}$ hashes requires more than 71 thousand years of computation time, and for $k = 3$ more than 1.5 trillion years of computation time. Given the scale of operations required, we consider a preimage attack as too costly for $k \geq 2$. This consideration is valid for interdisciplinary work, however attack methods which focus on a narrow, non-interdisciplinary field of research will be a number of magnitude cheaper.

5 CONCLUSION AND FUTURE WORK

We proposed BC-PSI - a method that computes the bibliographic coupling strength of documents without revealing the bibliographic references involved in the computation. To realize BC-PSI, we invented a new PSI approach that uses hashed feature subsets to prevent preimage and dictionary attacks. The security of the approach is adjustable to the computing resources that might be spent on the specific problem by increasing the number of features included in a subset, and by using hash functions with higher bit-lengths.

We showed that BC-PSI capably identifies similar documents in a large corpus without introducing hash collisions using SHA1. We demonstrated that a subset size of $k = 2$ achieves the best trade-off between computation time, required storage, and attack resistance. A subset size of $k > 2$ causes a steep rise in computation time and is therefore limited to documents with small numbers of references.

Hashed document feature subsets show promise for building a future decentralized PD service since accuracy would only decrease if a document pair does not contain at least two matching references. As shown in our prior work [5], an overlap of two references generally does not constitute a similarity that is significant enough to identify a document as suspicious of plagiarism.

In the future, we will allow encrypting the document IDs of unpublished documents. By doing so, the PDS can still detect that an input document overlaps with already existing unpublished documents in the distributed reference database. However, the PDS can no longer determine the number of documents with which the input document shares content. By using incentive mechanisms, the owner of the unpublished document is motivated to share the document with the PDS privately. To avoid false positives, authors can, e.g., submit previous versions of rejected grant proposals to prove that they were the authors of the earlier document.

In this initial study, we focused entirely on Bibliographic Coupling and excluded more sophisticated citation-based plagiarism detection methods like Greedy Citation Tiling and Longest Common Citation Sequences [5]. In the future, we will devise content protection methods that support pattern-based PD approaches that use mathematical features [15] and images [13]. We will analyze these detection methods and examine which features need to be masked and which features can be shared openly during the detection process without revealing any semantic information.

In summary, we successfully conducted the first step towards our vision of a decentralized content protecting plagiarism detection [7]. The findings of our initial study confirm our research direction of using hashed document features, such as references, in-text citations, images, and formulae, to devise such a service.

To ensure the reproducibility of our experiments, our data and code are available at *https://github.com/ag-gipp/20CppdData*

REFERENCES

[1] Lutz Bornmann and Rüdiger Mutz. 2015. Growth Rates of Modern Science: A Bibliometric Analysis Based on the Number of Publications and Cited References. *Journal of the Association for Information Science and Technology* 66, 11 (2015), 2215–2222. https://doi.org/10.1002/asi.23329

[2] Hao Chen, Kim Laine, and Peter Rindal. 2017. Fast Private Set Intersection from Homomorphic Encryption. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 1243–1255. https://doi.org/10.1145/3133956.3134061

[3] Lily Chen, Stephen Jordan, Yi-Kai Liu, Dustin Moody, Rene Peralta, Ray Perlner, and Daniel Smith-Tone. 2016. *Report on Post-Quantum Cryptography.* Technical Report 8105. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8105

[4] Graham Cormode and S. Muthukrishnan. 2005. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms* 55, 1 (2005), 58 – 75. https://doi.org/10.1016/j.jalgor.2003.12.001

[5] Bela Gipp. 2014. *Citation-based Plagiarism Detection - Detecting Disguised and Cross-language Plagiarism using Citation Pattern Analysis.* Springer Vieweg. https://doi.org/10.1007/978-3-658-06394-8

[6] O. Goldreich, S. Micali, and A. Wigderson. 1987. How to Play ANY Mental Game. In *Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing*. 218–229. https://doi.org/10.1145/28395.28420

[7] Cornelius Ihle. 2019. A Privacy-Preserving and Decentralized Approach for Plagiarism Detection. In *Proceedings of the Doctoral Consortium at ACM/IEEE Joint Conference on Digital Libraries.*

[8] Arif E Jinha. 2010. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing* 23, 3 (2010), 258–263. https://doi.org/doi.org/10.1087/20100308

[9] Rob Johnson, Anthony Watkinson, and Michael Mabe. 2018. *The STM Report: An overview of scientific and scholarly journal publishing.* Technical Report 5th ed. https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf

[10] M. M. Kessler. 1963. Bibliographic coupling between scientific papers. *American Documentation* 14, 1 (1963), 10–25. https://doi.org/10.1002/asi.5090140103

[11] Moxie Marlinspike. 2014. The Difficulty Of Private Contact Discovery. https://signal.org/blog/contact-discovery/

[12] T. C. Maxino and P. J. Koopman. 2009. The Effectiveness of Checksums for Embedded Control Networks. *IEEE Transactions on Dependable and Secure Computing* 6, 1 (Jan. 2009), 59–72. https://doi.org/10.1109/TDSC.2007.70216

[13] Norman Meuschke, Christopher Gondek, Daniel Seebacher, Corinna Breitinger, Daniel Keim, and Bela Gipp. 2018. An Adaptive Image-based Plagiarism Detection Approach. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries.* https://doi.org/10.1145/3197026.3197042

[14] Norman Meuschke, Vincent Stange, Moritz Schubotz, and Bela Gipp. 2018. HyPlag: A Hybrid Approach to Academic Plagiarism Detection. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR).* https://doi.org/10.1145/3209978.3210177

[15] Norman Meuschke, Vincent Stange, Moritz Schubotz, Michael Kramer, and Bela Gipp. 2019. Improving Academic Plagiarism Detection for STEM Documents by Analyzing Mathematical Content and Citations. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries.* https://doi.org/10.1109/JCDL.2019.00026

[16] Mummoorthy Murugesan, Wei Jiang, Chris Clifton, Luo Si, and Jaideep Vaidya. 2010. Efficient privacy-preserving similar document detection. *The VLDB Journal* 19, 4 (Aug. 2010), 457–475. https://doi.org/10.1007/s00778-009-0175-9

[17] Sadegh Riazi, Mojan Javaheripi, Siam Umar Hussain, and Farinaz Koushanfar. 2019. MPCircuits: Optimized Circuit Generation for Secure Multi-Party Computation. In *Proceedings of the IEEE International Symposium on Hardware Oriented Security and Trust (HOST).* 198–207. https://doi.org/10.1109/HST.2019.8740831

[18] Phillip Rogaway and Thomas Shrimpton. 2004. Cryptographic Hash-Function Basics: Definitions, Implications, and Separations for Preimage Resistance, Second-Preimage Resistance, and Collision Resistance. In *Proceedings of the International Workshop on Fast Software Encryption (FSE).*

[19] Nik Unger, Sahithi Thandra, and Ian Goldberg. 2016. Elxa: Scalable Privacy-Preserving Plagiarism Detection. In *Proceedings of the ACM Workshop on Privacy in the Electronic Society*. 153–164. https://doi.org/10.1145/2994620.2994633

[20] Debora Weber-Wulff. 2014. *False Feathers: A Perspective on Academic Plagiarism.* Springer. https://doi.org/10.1007/978-3-642-39961-9

[21] Debora Weber-Wulff. 2019. Plagiarism Detectors Are a Crutch, and a Problem. *Nature* 567, 7749 (2019), 435–435. https://doi.org/10.1038/d41586-019-00893-5